Unravelling the genetic basis of hereditary disorders by high-throughput exome sequencing strategies

Jazayeri, Omid

Chapter 1

General introduction

Since the discovery of the DNA double helix in 1953 [1], several milestones have marked the path of the field of genomic research. One of these was the introduction of Sanger sequencing in the late 1970s, which provided a novel and effective technique to analyze and improve our insight into the human genome [2], but the most significant was the completion of The Human Genome Project (HGP), conducted by many laboratories around the world, in 2001 [3].

In concert with the completion of the HGP, an exciting new era of sequencing has begun. Driven by the technical challenge posed by the HGP, techniques for DNA sequencing have changed dramatically and led to the advent of next-generation sequencing (NGS) methods. These methods rely on the massively parallel sequencing of very short DNA fragments [4]. Within a relatively short period of time, NGS technologies have revolutionized research on the human genome and also that of other organisms [5] in various fields of genomics including DNA sequencing, expression studies through RNA sequencing (RNA-Seq) and epigenetic investigations (ChIP-Seq) [6,7].

With respect to DNA sequencing in humans, NGS-based techniques can be applied to entire genomes or to select parts of the genome (exome sequencing or gene panels). The methods are categorized into three groups based on their sample preparation methodology. First, in Whole Genome Sequencing (WGS), the entire genome of individual humans, which is composed of 3.2 billion nucleotides, can be sequenced directly within a week [8]. Second, in Whole Exome Sequencing (WES) or gene panels in which capture-by-hybridization selectively enriches regions or genes of interest, followed by the sequencing procedure. WES allows sequencing all of ~20,000 human genes following enrichment of ~180,000 exons, in dozens of individuals, within a few days [8]. Based on a similar approach, dedicated disease-specific gene panels may be used for targeted sequencing of particular gene sets of interest. In the third method, amplicon enrichment, genes (or regions) of interest are enriched by a multiplex-PCR, the products of which are then subjected to the sequencing procedure. This method is frequently applied for cancer genetics to detect causative mutations in tumour samples derived from, e.g., breast cancer [9] and thyroid cancer [10] patients. A simplified overview of the application of NGS techniques to detect/identify disease causing mutations is depicted in Figure 1.

Importantly, introduction of NGS techniques provided unprecedented possibilities for genome analyses. These advances and the sharp decreases in the cost of sequencing technologies [12] have paved the way for widespread use of WES for disease-gene discovery. Miller syndrome was the first rare Mendelian disorder for which the causal gene was identified using WES [13].

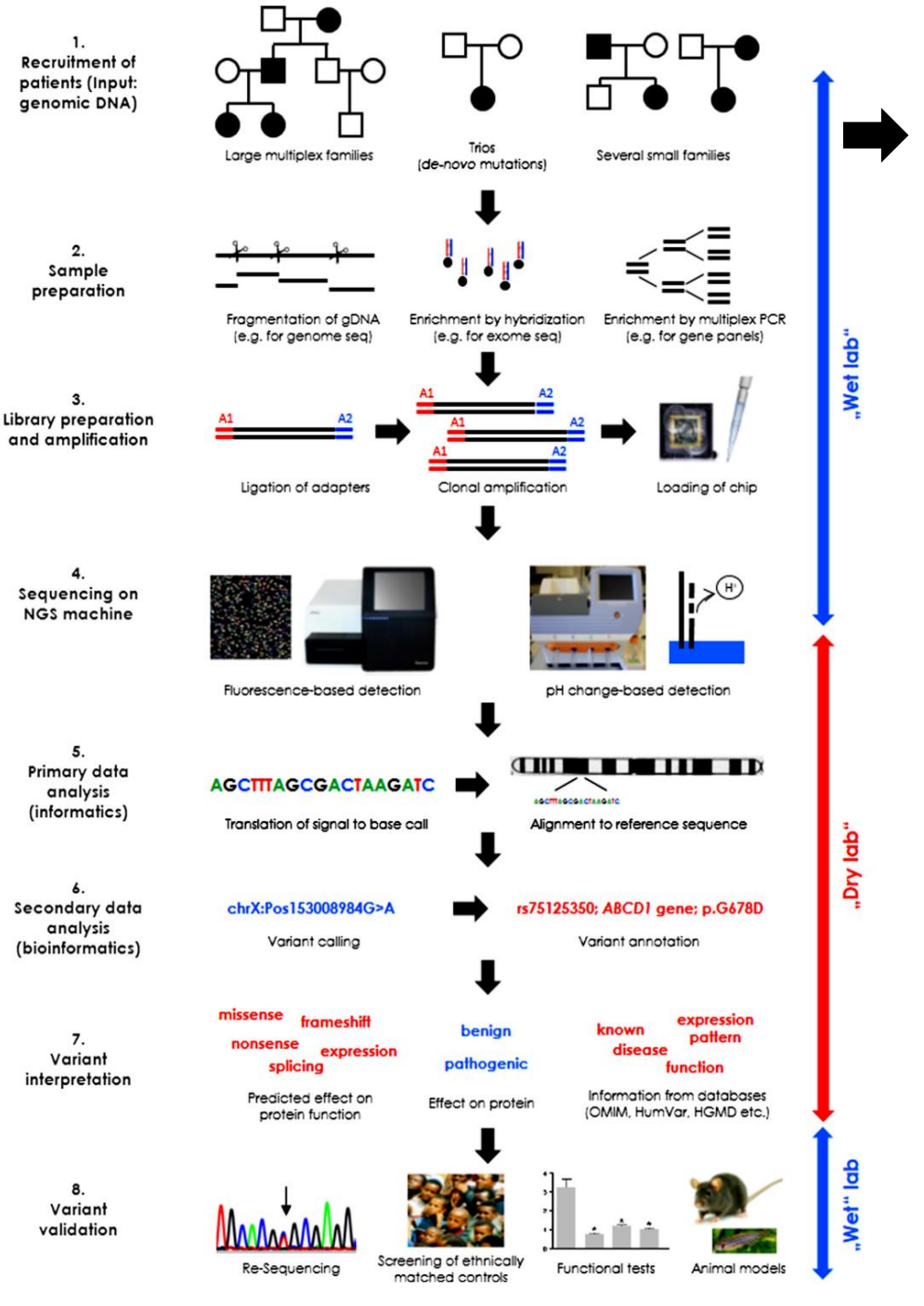
1. Recruitment of patients (Input: genomic DNA)
Large multiplex families
Trios (de-novo mutations)
Several small families
2. Sample preparation
Fragmentation of gDNA (e.g. for genome seq)
Enrichment by hybridization (e.g. for exome seq)
Enrichment by multiplex PCR (e.g. for gene panels)
3. Library preparation and amplification
A1
A2
Ligation of adapters
Clonal amplification
Loading of chip
4. Sequencing on NGS machine
Fluorescence-based detection
pH change-based detection
5. Primary data analysis (informatics)
AGCTTTAGCGACTAAGATC
Translation of signal to base call
Alignment to reference sequence
6. Secondary data analysis (bioinformatics)
chrX:Pos153008984G>A
Variant calling
rs75125350; ABCD1 gene; p.G678D
Variant annotation
7. Variant interpretation
missense
frameshift
nonsense
expression
splicing
benign
pathogenic
Predicted effect on protein function
Effect on protein
known disease
expression pattern
function
Information from databases (OMIM, HumVar, HGMD etc.)
8. Variant validation
Re-Sequencing
Screening of ethnically matched controls
Functional tests
Animal models
„Wet lab"
„Dry lab"
„Wet" lab

Figure 1: A simplified workflow of a next generation sequencing analysis (adapted from Lohmann & Klein [11]). **1.** For the sequencing analysis, DNA is required. This can be derived from large families with many affected patients, sporadic patients and their healthy parents (trios), cohorts of small families with the same disease (for gene discovery) or from individual patients or tissues (e.g. tumor samples). **2.** DNA is prepared for the sequencing by fragmentation. Specific target sequences need to be enriched (by hybridization or by PCR). **3.** Fragments are ligated to universal adapters and amplified clonally, and then loaded onto a chip. **4.** The sequencing reaction is monitored by a light signal (fluorescence) or by the release of a proton resulting in a pH change. **5.** The signals are translated into a sequence and aligned to the reference genome. **6.** Mismatches with the reference sequence are annotated with respect to the coding part of the genome. **7.** Distinction are made between likely benign or possibly pathogenic variants using *In silico* methods such as predictor algorithms or mutation databases which comprehensively collect pathogenic mutations. **8.** Finally, candidate variants classified as potential disease-causing need to be validated by Sanger sequencing along with functional studies and knockout animal models.

Since then, the list of rare Mendelian disorders for which WES was used to identify the genetic causes is rapidly growing [14]. WES has also been valuable in identifying causal and predisposing variants in common diseases and complex traits such as cardiovascular diseases, obesity and diabetes, hypertension and cancers [15]. The rationale for applying WES is evident: the 20,687 protein-coding genes [16] constitute only ~1% of the human genome but harbor about 85% of the known disease-associated mutations. In fact, most Mendelian disorders are caused by exonic mutations or splice-site mutations [17]. Accordingly, by sequencing no more than about 50Mb of our genome, the entire exome can be effectively analyzed. In recent years, WES has become the routine diagnostic application of NGS to identify the causes of mostly monogenic disorders. In particular, exome sequencing has helped to overcome difficulties and limitations in the identification of causal mutations in diseases with extreme locus heterogeneity such as intellectual disabilities [18], hearing loss [19] and retinitis pigmentosa [20].

One of the unique capabilities NGS methods offer is the detection of *de novo* variants across the entire genome by application of a trio-based sequencing strategy. This has unraveled the genetic cause, including identification of novel disease-associated genes, in sporadic cases of autism [21], intellectual disability [22] and schizophrenia [23].

Although WES may lead to identification of causal mutation(s), and despite its unique ability to detect *de novo* variants genome-wide, limitations of the technique such as incomplete capturing of all exons and the complexity of variant interpretation still pose challenges for application of WES in the clinic [24–26].

In parallel to applying WES, a broad range of dedicated gene panels have been implemented in germ-line and tumor diagnostics. These provide a sensitivity and specificity that equals the current standard of Sanger sequencing [27] and, to

date, outperform WES in this respect. Comprehensive gene panels also allow more samples to be analyzed in a single run (making this technique cost-efficient) with a shorter turn-around time than clinical exome sequencing [28–30]. Gene panels have been implemented to diagnose mitochondrial disorders [31], epilepsy [32], retinal dystrophies [29], inflammatory bowel disease [33], cardiomyopathies [34] as well as amplicon-based targeted sequencing of breast cancer [9] and myeloid malignancies [35]. In addition to the reliable detection of mutations, such panels are able to detect even low-percentage mosaicisms like those seen in tumor materials [28,36].

Given that sequencing of the entire human genome has not been a generally accessible option until recently, the first real success of WGS was that it became possible to accurately detect and measure fetal DNA in maternal serum to screen for fetal aneuploidies using a noninvasive method [37–39]. More recently, however, with the availability of the newer generation of high-capacity sequencers, together with a further reduction in sequencing costs, WGS is being considered more and more often as the ultimate option for diagnostic testing. In addition to exomic variants, WGS is able to detect pathogenic structural variations like inter- and intrachromosomal translocation or chromosomal rearrangements [40–42].

With all these exiting possibilities at hand, and further technological improvements expected, the biggest challenge in using NGS techniques lies in the data-interpretation. For instance, exome sequencing yields approximately between 20,000 and 50,000 variants per sequenced exome. These figures vary enormously among different sequencing studies depending on the ethnicity of the patients [43], the method used for exome enrichment, the sequencing platform, sequencing depth, and the algorithms used for mapping and variant calling [44]. Identifying causal mutation(s) against a huge background of non-pathogenic polymorphisms and sequencing errors is a key challenge when interpreting NGS results. For instance, each human genome harbors about 100 genuine loss-of-function variants with around 20 genes completely inactivated [45]. Thus, having as much complementary information as possible, such as the pedigree or population structure, inheritance pattern, phenotypic features, known or unknown etiology, knowing whether a phenotype arises owing to *de novo* or inherited variants, linkage and/or homozygosity mapping data, is of utmost importance for optimal customized variant interpretation and prioritization strategies.

Aim of this thesis

The research presented in this thesis focuses on WES applications for unravelling the genetic basis of human hereditary disorders by (i) screening of a comprehensive set of known disease-causing genes for mutations; (ii) identifying novel candidate disease genes in genetically heterogeneous disorders with different possible inheritance patterns; and (iii) developing bioinformatic tools to improve on the possibilities for selecting putative candidate genes by *in silico* analyses.

A diagram summarizing the structure of this thesis is shown in Figure 2. In **Chapter 2**, I examine the utility of WES as a diagnostic approach for establishing a molecular diagnosis in a highly heterogeneous group of patients with microcephaly and variable intellectual disability. Many patients with microcephaly and intellectual disability have a rather indistinguishable and non-specific phenotype. Diagnostic testing through candidate gene testing by Sanger sequencing is rather limited and the diagnostic yield is generally low. We therefore set out to apply WES as a tool to survey all microcephaly genes reported thus far in a single test and to determine WES's usefulness in diagnosing patients.

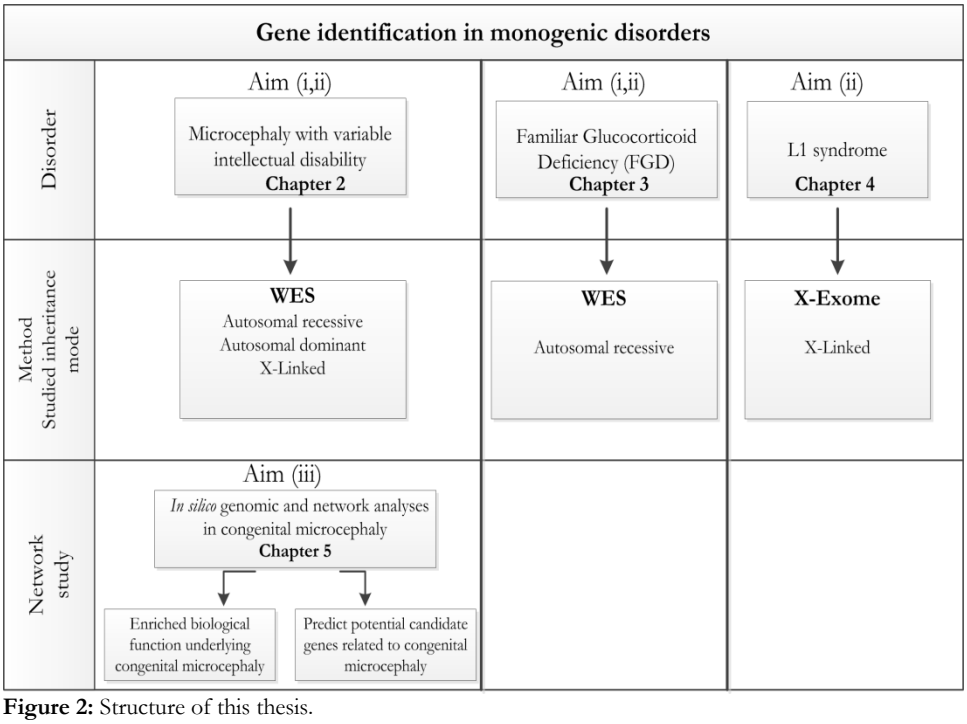

Figure 2: Structure of this thesis.

In **Chapter 3** we set out to identify the disease gene in a Dutch family with a boy affected by familiar glucocorticoid deficiency (FGD) by applying trio-exome sequencing.

Chapter 4 aims to identify a novel disease causing gene for L1 syndrome, an X-linked disorder, by application of a variant of WES. To date, L1 syndrome is known to be caused by mutations in the *L1CAM* gene, but many unsolved cases remain after mutation analysis of this gene. Pursuing the hypothesis that the disorder is considered to be X-linked, i.e. that the causative gene should be located on the X chromosome, we initiated X-exome sequencing in a cohort of male patients.

Chapter 5 describes the development of an *in silico* composite network to predict novel candidate genes that might be involved in congenital microcephaly. Unsolved cases are a prevalent problem in diagnostics and researchers are often confronted with the limitations of WES to detect pathogenic mutations. Non-coding variants or inadequate quality of sequence in true causative gene may lead investigators to miss causal mutations. In addition, and importantly, undiscovered gene(s) associated with a given disorder are clearly a main explanation for unsolved cases. In the latter situation, applying *in silico* methods to create composite biological network from publically available database-resources could provide new, additional candidate genes to be considered in identifying disease-genes.

Finally in **Chapter 6** I will discuss the effectiveness and limitations of applying WES for molecular genetic studies on human disease, as well as discussing possible improvements and presenting my future perspectives on high-throughput sequencing approaches.

References

1. Watson JD, Crick FH, others (1953) Molecular structure of nucleic acids. Nature 171: 737–738.
2. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences 74: 5463–5467.
3. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. (2001) The sequence of the human genome. Science 291: 1304–1351.
4. Shendure J, Ji H (2008) Next-generation DNA sequencing. Nature biotechnology 26: 1135–1145.
5. Ekblom R, Galindo J (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity (Edinb) 107: 1–15.
6. Fouse SD, Nagarajan RP, Costello JF (2010) Genome-scale DNA methylation analysis. Epigenomics 2: 105–117.
7. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics 10: 57–63.
8. Marian AJ (2012) Challenges in medical applications of whole exome/genome sequencing discoveries. Trends in cardiovascular medicine 22: 219–223.
9. Tung N, Battelli C, Allen B, Kaldate R, Bhatnagar S, et al. (2015) Frequency of mutations in individuals with breast cancer referred for BRCA1 and BRCA2 testing using next-generation sequencing with a 25-gene panel. Cancer 121: 25–33.
10. Nikiforova MN, Wald AI, Roy S, Durso MB, Nikiforov YE (2013) Targeted next-generation sequencing panel (ThyroSeq) for detection of mutations in thyroid cancer. The Journal of Clinical Endocrinology & Metabolism 98: 1852–1860.
11. Lohmann K, Klein C (2014) Next Generation Sequencing and the Future of Genetic Diagnosis. Neurotherapeutics 11: 699–707.
12. Sboner A, Mu XJ, Greenbaum D, Auerbach RK, Gerstein MB, et al. (2011) The real cost of sequencing: higher than you think! Genome Biology 12: 125.
13. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, et al. (2010) Exome sequencing identifies the cause of a mendelian disorder. Nature genetics 42: 30–35.
14. Rabbani B, Mahdieh N, Hosomichi K, Nakaoka H, Inoue I (2012) Next-generation sequencing: impact of exome sequencing in characterizing Mendelian disorders. Journal of human genetics 57: 621–632.
15. Rabbani B, Tekin M, Mahdieh N (2014) The promise of whole-exome sequencing in medical genetics. Journal of human genetics 59: 5–15.
16. Pennisi E (2012) ENCODE project writes eulogy for junk DNA. Science 337: 1159–1161.
17. Majewski J, Schwartzentruber J, Lalonde E, Montpetit A, Jabado N (2011) What can exome sequencing do for you? Journal of medical genetics 48: 580–589.

18. Hamdan FF, Srour M, Capo-Chichi J-M, Daoud H, Nassif C, et al. (2014) De Novo Mutations in Moderate or Severe Intellectual Disability. PLoS genetics 10: e1004772.
19. McClellan J, King M-C (2010) Genetic heterogeneity in human disease. Cell 141: 210–217.
20. Daiger SP, Sullivan LS, Bowne SJ (2013) Genes and mutations causing retinitis pigmentosa. Clinical genetics 84: 132–141.
21. O'Roak BJ, Deriziotis P, Lee C, Vives L, Schwartz JJ, et al. (2011) Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. Nature genetics 43: 585–589.
22. De Ligt J, Willemsen MH, van Bon BW, Kleefstra T, Yntema HG, et al. (2012) Diagnostic exome sequencing in persons with severe intellectual disability. New England Journal of Medicine 367: 1921–1929.
23. Rees E, Kirov G, O'Donovan MC, Owen MJ (2012) De novo mutation in schizophrenia. Schizophrenia bulletin 38: 377–381.
24. Xuan J, Yu Y, Qing T, Guo L, Shi L (2013) Next-generation sequencing in the clinic: promises and challenges. Cancer letters 340: 284–295.
25. De Leeneer K, Hellemans J, De Schrijver J, Baetens M, Poppe B, et al. (2011) Massive parallel amplicon sequencing of the breast cancer genes BRCA1 and BRCA2: opportunities, challenges, and limitations. Human mutation 32: 335–344.
26. Frebourg T (2014) The Challenge for the Next Generation of Medical Geneticists. Human mutation 35: 909–911.
27. Sikkema-Raddatz B, Johansson LF, Boer EN, Almomani R, Boven LG, et al. (2013) Targeted Next-Generation Sequencing can Replace Sanger Sequencing in Clinical Diagnostics. Human mutation 34: 1035–1042.
28. Chong HK, Wang T, Lu H-M, Seidler S, Lu H, et al. (2014) The Validation and Clinical Implementation of BRCAplus: A Comprehensive High-Risk Breast Cancer Diagnostic Assay. PLoS One 9: e97408.
29. Glöckle N, Kohl S, Mohr J, Scheurenbrand T, Sprecher A, et al. (2013) Panel-based next generation sequencing as a reliable and efficient technique to detect mutations in unselected patients with retinal dystrophies. European Journal of Human Genetics 22: 99–104.
30. Group SM (2015) Comprehensive gene panels provide advantages over clinical exome sequencing for Mendelian diseases. Genome biology 16: 134.
31. Platt J, Cox R, Enns GM (2014) Points to Consider in the Clinical Use of NGS Panels for Mitochondrial Disease: An Analysis of Gene Inclusion and Consent Forms. Journal of genetic counselling 23: 594–603.
32. Wang J, Gotway G, Pascual JM, Park JY (2014) Diagnostic Yield of Clinical Next-Generation Sequencing Panels for Epilepsy. JAMA neurology 71: 650–651.
33. Kammermeier J, Drury S, James CT, Dziubak R, Ocaka L, et al. (2014) Targeted gene panel sequencing in children with very early onset inflammatory bowel disease—evaluation and prospective analysis. Journal of medical genetics 51: 748–755.

34. D'Argenio V, Frisso G, Precone V, Boccia A, Fienga A, et al. (2014) DNA sequence capture and next-generation sequencing for the molecular diagnosis of genetic cardiomyopathies. The Journal of molecular diagnostics 16: 32–44.
35. Cheng DT, Cheng J, Mitchell TN, Syed A, Zehir A, et al. (2014) Detection of mutations in myeloid malignancies through paired-sample analysis of microdroplet-PCR deep sequencing data. The Journal of Molecular Diagnostics 16: 504–518.
36. Friedman E, Efrat N, Soussan-Gutman L, Dvir A, Kaplan Y, et al. (2015) Low-level constitutional mosaicism of a de novo BRCA1 gene mutation. British Journal of Cancer 112: 765–768.
37. Lo YD (2013) Non-invasive prenatal testing using massively parallel sequencing of maternal plasma DNA: from molecular karyotyping to fetal whole-genome sequencing. Reprod Biomed Online 27: 593–598.
38. Lo YD, Chan KA, Sun H, Chen EZ, Jiang P, et al. (2010) Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. Science translational medicine 2: 61ra91.
39. Chen EZ, Chiu RWK, Sun H, Akolekar R, Chan KCA, et al. (2011) Noninvasive prenatal diagnosis of fetal trisomy 18 and trisomy 13 by maternal plasma DNA sequencing. PLoS One 6: e21791.
40. Nishiguchi KM, Tearle RG, Liu YP, Oh EC, Miyake N, et al. (2013) Whole genome sequencing in patients with retinitis pigmentosa reveals pathogenic DNA structural changes and NEK2 as a new disease gene. Proceedings of the National Academy of Sciences 110: 16139–16144.
41. Gilissen C, Hehir-Kwa JY, Thung DT, van de Vorst M, van Bon BW, et al. (2014) Genome sequencing identifies major causes of severe intellectual disability. Nature 511: 344–348.
42. Wei L, Liu S, Conroy J, Wang J, Papanicolau-Sengos A, et al. (2015) Whole-genome sequencing of a malignant granular cell tumor with metabolic response to pazopanib. Molecular Case Studies 1: a000380.
43. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, et al. (2011) Exome sequencing as a tool for Mendelian disease gene discovery. Nature Reviews Genetics 12: 745–755.
44. Gilissen C, Hoischen A, Brunner HG, Veltman JA (2012) Disease gene identification strategies for exome sequencing. European Journal of Human Genetics 20: 490–497.
45. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, et al. (2012) A systematic survey of loss-of-function variants in human protein-coding genes. Science 335: 823–828.